Appendix

Appendix A1.1   Study characteristics: Baker, 1997 (randomized controlled trial)

Characteristic	Description
Study citation	Baker, J. J. (1997). Effects of a generative instructional design strategy on learning mathematics and on attitudes towards achievement. *Dissertation Abstracts International, 58*(7), 2573A. (UMI No, 9800955).
Participants	Participants were 90 eighth-grade students. Most students were from low-income families and qualified for the free or reduced-price lunch program. All but three students were white, and none were in special education. Students were randomly assigned to the intervention or comparison condition; however, some participants left the school before posttest, some enrolled after pretest, and some missed either the pretest or posttest. Seventy students (34 intervention and 36 comparison) completed both the pre- and posttests and were included in the statistical analysis.
Setting	The study took place in a suburban middle school in St. Louis, Missouri. There were four classrooms—two intervention and two comparison.
Intervention	The intervention group used *Transition Mathematics*, first edition. The study referred to this curriculum as a "linear curriculum," which is teacher-directed and uses fixed media. This mode of teaching was contrasted with a "generative curriculum," which was described as more self-generated (with the guidance of printed materials) but teacher illuminated. Students worked individually or in groups using the curriculum throughout the school year. The same teacher taught both the intervention and the comparison groups. Class periods were block scheduled, about 85 minutes long, covering 1–2 lessons. Homework was assigned almost every class and was sometimes completed by students during the class.
Comparison	The comparison group was taught using a "generative mathematics curriculum" that used the *Expert Mathematician,* version 3.0. Students worked individually or in pairs using printed materials and the computer to complete lessons in the *Expert Mathematician.* Sessions, which included one or two lessons, were 85 minutes long and occurred every other day for one school year. Comparison students were taught in the same classroom by the same teacher as those in the intervention condition. Homework was not assigned in every class.
Primary outcomes and measurement	The Objectives by Strand test was administered at the end of the school year by the classroom teacher. (See Appendix A2 for more detailed descriptions of outcome measures.)
Teacher training	No information about teacher training was provided.

Appendix A1.2   Study characteristics: Hedges, Stodolsky, Mathison, & Flores, 1986 (quasi-experimental design)

Characteristic	Description
Study citation	Hedges, L. V., Stodolsky, S. S., Mathison, S., & Flores, P. V. (1986). *Transition Math Field Study.* Chicago: University of Chicago.
Participants	The study included seventh-grade students in the 50th-90th percentile of math ability, eighth-grade students in the 30th-70th percentile, and ninth-grade students in the 15th-50th percentile based on standardized test scores. The participants were enrolled in a pre-algebra class. The original sample (referred to in the study as "all pairs") included 1,048 students in 41 intervention classrooms and 976 students in 38 comparison classrooms. Based on examination of the statistical significance of pretest differences and class size and composition, the study defined a subset of 20 well-matched pairs of classrooms (7 pairs of seventh grade, 10 pairs of eighth grade, and 3 pairs of ninth grade). This sample was referred to in the study as "the conservative sample." The sample was diverse in terms of demographic characteristics—for example, socioeconomic status and ethnic minority.
Setting	The "all pairs" sample was recruited from 35 schools in rural, suburban, and urban areas in 10 states: Colorado, Florida, Illinois (Chicago and Chicago suburbs), Maryland, Michigan, Mississippi, New Mexico, Ohio, Virginia, and Washington. No information was provided on the number of schools and geographical areas from which the conservative sample was recruited.
Intervention	This study used the first edition of the *Transition Mathematics* textbook, which addresses pre-algebra, pre-geometry, and applied arithmetic. Compared with traditional curricula, *Transition Mathematics* emphasizes reading math and using calculators. Lessons ranged 40–60 minutes across grade levels. Teachers were expected to do a lesson a day, but the study reported variations in the number of lessons taught each week. The study reported that lower student math ability made it difficult to implement the intervention as intended.
Comparison	The comparison group was taught using curricula that were not *Transition Mathematics* but provided similar opportunities to learn mathematics content.
Primary outcomes and measurement	Primary outcomes used in this study were the High School Subjects Test: General Mathematics, the Geometry Readiness test, and the Orleans-Hanna Algebra Prognosis test. An additional nonstandardized measure, the Alternative Test, was not reviewed because complete statistical information was not available, so effect sizes could not be computed for this measure. (See Appendix A2 for more detailed descriptions of outcome measures.)
Teacher training	No information about teacher training was provided.

Appendix A1.3   Study characteristics: Thompson, Senk, Witonsky, Usiskin, & Kaeley, 2005 (quasi-experimental design)

Characteristic	Description
Study citation	Thompson, D. R., Senk, S. L., Witonsky, D., Usiskin, Z., & Kaeley, G. (2005). *An evaluation of the second edition of UCSMP Transition Mathematics.* Chicago: University of Chicago School Mathematics Project.
Participants	The study used a matched pair design in which classrooms of students in the same school were matched based on students' mathematical ability. One school included two matched pairs of classrooms, and two schools each included one matched pair of classrooms, for a total of eight participating classrooms. The study included 91 students (41 intervention and 50 comparison) enrolled in grades 7–9. The percentage of female students in the intervention group was 44% and 50% in the comparison group. The study reported that no information on ethnicity was solicited from the students.
Setting	The study included an inner-city junior high school in the West, a rural middle school in the Southeast, and an inner-city high school in the upper Midwest. Intervention and comparison classrooms were drawn from each of the schools. The three schools served a diverse student population that included low socioeconomic status minority populations.
Intervention	The intervention group used the *Transition Mathematics* second edition textbook, which was tested in this study and revised afterwards. Teachers were expected to complete 10–12 of the 13 chapters in the textbook, with a median of 11 chapters. In one school, however, teachers were expected to cover only six chapters. Time spent on instruction during class ranged 40–50 minutes. About 25% of students had little or no access to calculators, and some of the intervention classrooms did not have access to computers.
Comparison[1]	The study reported that comparison classrooms used traditional materials currently in use in the schools. No information was solicited from teachers about the number of chapters they were expected to cover. Time spent on instruction during class ranged 40–50 minutes. About 46% of students had little or no access to calculators, and some of the comparison classrooms did not have access to computers.
Primary outcomes and measurement	This study used four measures of mathematics achievement—High School Subjects Test: General Mathematics, Geometry Readiness test, Algebra Readiness test, and Problem-Solving and Understanding test. Achievement on the High School Subjects Test: General Mathematics and the Algebra Readiness test was analyzed in three ways: overall achievement using all test items; using only the test items for which all study participants in the same school had studied the content needed to answer the items, so the test was different in each school (referred to as "fair test"); and using only the test items for which all students in all the participating schools had studied the content needed to answer the items, so the test was the same in all schools (referred to as "conservative test"). For this WWC review, only the analyses focusing on overall achievement were used for rating purposes. (See Appendix A2 for more detailed descriptions of outcome measures.)
Teacher training	No information was provided on teacher training.

1. Thompson et al. (2005) also compared *Transition Mathematics,* second edition with *Transition Mathematics,* first edition and found no statistically significant differences.

Appendix A2   Outcome measures in the mathematics achievement domain

Outcome measure	Description
Objectives by Strand	A 78-item multiple-choice test of mathematical ability (as cited in Baker, 1997).
Orleans-Hanna Algebra Prognosis test	This nationally normed test consists of 60 multiple-choice items based on nine model lessons and five questionnaire items that require students to report their course grades and predict their final grade if they were to take algebra. In contrast to an achievement test, students are required to answer questions by following a procedure or set of operations using mathematical or verbal expressions parallel to but different from those contained in the model lessons. This test is often used to predict the ability to succeed in a first-year algebra course (as cited in Hedges et al., 1986).
High School Subjects Test: General Mathematics	This test consists of 40 items—20 arithmetic skills items and 20 arithmetic uses items. The test was developed by Scott, Foresman, and Company (as cited in Hedges et al., 1986; Thompson et al., 2005).
Geometry Readiness	This 19-item test was constructed by UCSMP and adapted from a test developed in the early 1970s for the Cognitive Development and Achievement in the Secondary School Geometry Study (as cited in Hedges et al., 1986; Thompson et al., 2005). According to the study's description of the test, it covers relevant topics in geometry, including properties of lines; perimeter, area, and volume; figures and their properties; angles and their measures; and circles. The test was accepted for review based on the description of its content.
Algebra Readiness	This test consists of 11 multiple-choice items and 10 short constructed-response items. The test was developed by UCSMP (as cited in Thompson et al., 2005). According to the study's description of the test, it covers relevant topics in algebra, including translating words into symbols, properties of numbers, scientific notations, graphs of lines, equation solving, coordinates, estimation, and evaluating an expression. The test was accepted for review based on the description of its content.
Problem-Solving and Understanding	This open-ended problem-solving test was developed at the UCSMP (as cited in Thompson et al., 2005). This test was administered in two different forms—"even form" and "odd form." Each form examines different mathematics skills and knowledge (for example, probability, mean and median, decimals and fractions, and area and perimeter). Half of the students in each class were randomly assigned to each form. The test was accepted for review based on the description of its content.

Outcome measure	Study sample	Sample size (schools/ students)	Author's findings from the study		Mean difference[4] (*Transition Mathematics* – comparison)	WWC calculations		
			Mean outcome (standard deviation[2])					
			Transition Mathematics group[3]	Comparison group		Effect size[5]	Statistical significance[6] (at $\alpha = 0.05$)	Improvement index[7]
Baker, 1997 (randomized controlled trial)[8]								
Objectives by Strand	Grade 8	1/70	40.8 (12.41)	45.1 (12.03)	−4.30[9]	−0.35	ns	−14
Average[10] for mathematics achievement (Baker, 1997)						−0.35	ns	−14
Hedges et al., 1986 (quasi-experimental design)[8]								
Orleans-Hanna Algebra Prognosis test	Grades 7–9	40/809	38.91 (13.22)	36.66 (13.67)	2.24	0.17	ns	+7
Geometry Readiness	Grades 7–9	40/587	9.87 (4.17)	8.67 (4.15)	1.20	0.29	Statistically significant	+11
High School Subjects Test: General Mathematics	Grades 7–9	40/806	26.17 (8.00)	25.12 (7.60)	1.05	0.13	ns	+5
Average[10] for mathematics achievement (Hedges et al., 1986)						0.20	ns	+8
Thompson et al., 2005 (quasi-experimental design)[8]								
High School Subjects Test: General Mathematics	Grades 7–9	8/91	52.70 (23.90)	58.5 (20.60)	−5.80	−0.26	ns	−10
Algebra Readiness	Grades 7–9	8/91	41.90 (25.10)	39.7 (22.30)	2.20	0.09	ns	+4
Geometry Readiness	Grades 7–9	8/91	49.20 (19.10)	39.5 (18.90)	9.70	0.51	ns	+19
Problem-Solving and Understanding	Grades 7–9	8/93	4.02 (2.99)[11]	2.96 (3.06)	1.06	0.35	ns	+14
Average[10] for mathematics achievement (Thompson et al., 2005)						0.17	ns	+7
Domain average[10] for mathematics achievement across all studies						0.02	na	+0

ns = not statistically significant
na = not applicable

(continued)

Appendix A3   Summary of study findings included in the rating for the mathematics achievement domain

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices. Subgroup findings are presented in appendix A4.1. Findings by form type (even form and odd form) of High School Subjects Test: General Mathematics in Thompson et al. (2005) are presented in Appendix A4.2.
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes. Standard deviations and means for Hedges et al. (1986) were requested by the WWC and received from the first study author.
3. In Baker (1997), the intervention group is *Expert Mathematician,* but in this report the intervention group is *Transition Mathematics.*
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
5. For an explanation of the effect size calculation, see Technical Details of WWC-Conducted Computations.
6. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
7. The improvement index represents the difference between the percentile rank of the average student in the intervention condition and that of the average student in the comparison condition. The improvement index can take on values between −50 and +50, with positive numbers denoting results favorable to the intervention group.
8. The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. See Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate statistical significance. In the case of Baker (1997), no corrections for clustering or multiple comparisons were needed. In the case of Hedges et al. (1986), the study reported that the analyses were conducted on the classroom level, so corrections for clustering were not needed; however, corrections for multiple comparisons were needed. In the case of Thompson et al. (2005), corrections for clustering and multiple comparisons were needed.
9. In Baker (1997), the mean difference was calculated using the difference in difference approach, which takes into account baseline differences; in this table, the comparison group mean is the posttest intervention group mean plus the absolute value of the mean difference.
10. The WWC-computed average effect sizes for each study and for the domain across studies are simple averages rounded to two decimal places. The average improvement indices are calculated from the average effect size.
11. Means and standard deviations for the Problem-Solving and Understanding test were calculated by the WWC based on the means and standard deviations presented for the even and odd forms of this test, taking into account the number of students who completed each type of form (see Appendix A4.2).

Appendix A4.1   Summary of subgroup findings for the mathematics achievement domain[1]

Outcome measure	Study sample	Sample size (schools/ students)	Author's findings from the study		WWC calculations			
			Mean outcome (standard deviation[2])					
			Transition Mathematics group	Comparison group	Mean difference[3] (Transition Mathematics – comparison)	Effect size[4]	Statistical significance[5] (at $\alpha = 0.05$)	Improvement index[6]
Hedges et al., 1986 (quasi-experimental design)[7]								
High School Subjects Test: General Mathematics	Grades 7–9 (using calculators)	40/232	29.86 (6.55)	26.46 (7.20)	3.39	0.49	Statistically significant	+19
High School Subjects Test: General Mathematics	Grades 7–9 (without calculators)[8]	40/574	24.59 (8.06)	24.61 (7.70)	−0.02	0.00	ns	+0

ns = not statistically significant

1. This appendix presents subgroup findings for measures that fall in the mathematics achievement domain. Total scores for the entire sample were used for rating purposes and are presented in Appendix A3.

2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes. For Hedges et al. (1986), the WWC requested and received from the first study author posttest standard deviations.

3. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.

4. For an explanation of the effect size calculation, see Technical Details of WWC-Conducted Computations.

5. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.

6. The improvement index represents the difference between the percentile rank of the average student in the intervention condition and that of the average student in the comparison condition. The improvement index can take on values between −50 and +50, with positive numbers denoting results favorable to the intervention group.

7. The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools (corrections for multiple comparisons were not done for findings not included in the overall intervention rating). For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. See Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate statistical significance. In the case of Hedges et al. (1986), no correction for clustering was needed.

8. Students without calculators did not have personal calculators and were not provided calculators by the school.

Appendix A4.2   Summary of additional findings for the mathematics achievement domain[1]

Outcome measure	Study sample	Sample size (schools/ students)	Author's findings from the study		WWC calculations			
			Mean outcome (standard deviation[2])					
			Transition Mathematics group	Comparison group	Mean difference[3] (Transition Mathematics – comparison)	Effect size[4]	Statistical significance[5] (at α = 0.05)	Improvement index[6]
Thompson et al., 2005 (quasi-experimental design)[7]								
Problem-Solving and Understanding test: odd form	Grades 7–9	4/47	3.95 (2.48)	3.72 (3.34)	0.23	0.08	ns	+3
Problem-Solving and Understanding test: even form	Grades 7–9	4/46	4.10 (3.50)	2.20 (2.60)	1.90	0.61	ns	+23

ns = not statistically significant

1. This appendix presents findings for the even form and odd form versions of the Problem-Solving and Understanding test reported in Thompson et al. (2005). Aggregated scores across forms were used for rating purposes and are presented in Appendix A3.
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
4. For an explanation of the effect size calculation, see Technical Details of WWC-Conducted Computations.
5. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
6. The improvement index represents the difference between the percentile rank of the average student in the intervention condition and that of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
7. The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools (corrections for multiple comparisons were not done for findings not included in the overall intervention rating). For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. See Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate statistical significance. In the case of Thompson et al. (2005), correction for clustering was needed.

Appendix A5 *Transition Mathematics* rating for the mathematics achievement domain

The WWC rates an intervention's effects in a given outcome domain as positive, potentially positive, mixed, no discernible effects, potentially negative, or negative.[1]

For the outcome domain of mathematics achievement, the WWC rated *Transition Mathematics* as having mixed effects. It did not meet the criteria for positive or potentially positive effects because one study showed a substantively important negative effect. Further, it did not meet the criteria for negative or potentially negative effects because one study showed a statistically significant positive effect. The remaining rating (no discernible effects) was not considered because *Transition Mathematics* was assigned the highest applicable rating.

Rating received

Mixed effects: Evidence of inconsistent effects as demonstrated through either of the following criteria.

- Criterion 1: At least one study showing a statistically significant or substantively important *positive* effect, and at least one study showing a statistically significant or substantively important *negative* effect, but no more such studies than the number showing a statistically significant or substantively important *positive* effect.

  Met. One study showed a substantively important negative effect, and one study showed a statistically significant positive effect. One study showed indeterminate effects.

- Criterion 2: At least one study showing a statistically significant or substantively important effect, and more studies showing an *indeterminate* effect than showing a statistically significant or substantively important effect.

  Not met. One study showed a substantively important negative effect, and one study showed a statistically significant positive effect. Only one study showed indeterminate effects.

Other ratings considered

Positive effects: Strong evidence of a positive effect with no overriding contrary evidence.

- Criterion 1: Two or more studies showing statistically significant *positive* effects, at least one of which met WWC evidence standards for a strong design.

  Not met. The one study that showed a statistically significant positive effect did not use a strong design.

- Criterion 2: No studies showing statistically significant or substantively important *negative* effects.

  Not met. One study showed a substantively important negative effect.

Potentially positive effects: Evidence of a positive effect with no overriding contrary evidence.

- Criterion 1: At least one study showing a statistically significant or substantively important *positive* effect.

  Met. One study that met WWC standards with reservations showed a statistically significant positive effect.

- Criterion 2: No studies showing a statistically significant or substantively important *negative* effect and fewer or the same number of studies showing *indeterminate* effects than showing statistically significant or substantively important *positive* effects.

  Not met. One study showed a substantively important negative effect.

(continued)

Potentially negative effects: Evidence of a negative effect with no overriding contrary evidence

- Criterion 1: At least one study showing a statistically significant or substantively important *negative* effect.

    Met. One study showed a substantively important negative effect.

- Criterion 2: No studies showing a statistically significant or substantively important *positive* effect, or more studies showing statistically significant or substantively important *negative* effects than showing statistically significant or substantively important *positive* effects.

    Not met. One study showed a statistically significant positive effect.

Negative effects: Strong evidence of a negative effect with no overriding contrary evidence.

- Criterion 1: Two or more studies showing statistically significant *negative* effects, at least one of which met WWC evidence standards for a strong design.

    Not met. No studies showed a statistically significant negative effect, and no studies met WWC evidence standards for a strong design.

- Criterion 2: No studies showing statistically significant or substantively important *positive* effects.

    Not met. One study showed a statistically significant positive effect.

1. For rating purposes, the WWC considers the statistical significance of individual outcomes and the domain level effect. The WWC also considers the size of the domain level effect for ratings of potentially positive or potentially negative effects. See the WWC Intervention Rating Scheme for a complete description.